AFRL-RH-WP-TR-2010-0008

Factor Structure of the Air Force Officer Qualifying Test Form S: Analysis and Comparison with Previous Forms

Fritz Drasgow
Christopher D. Nye
University of Illinois at Urbana-Champaign
603 E. Daniel St.
Champaign IL 61820

Thomas R. Carretta
Warfighter Interface Division
Supervisory Control Interfaces Branch

Malcolm James Ree
Our Lady of the Lake University
San Antonio TX 78207

October 2008
Interim Report for May 2008 to October 2008

Approved for public release;
Distribution is unlimited.

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Supervisory Control Interfaces Branch
Wright-Patterson AFB OH 45433

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th ABW Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2010-0008 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR

//**signed//**
Thomas R. Carretta
Research Engineering Psychologist
Supervisory Control Interfaces Branch

//signed//
Daniel G. Goddard
Chief, Warfighter Interfaces Division
Human Effectiveness Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government’s approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE	*Form Approved* **OMB No. 0704-0188**

1. REPORT DATE *(DD-MM-YYYY)* 17 Oct 2008	**2. REPORT TYPE** Interim	**3. DATES COVERED** *(From - To)* May 2008 – October 2008

4. TITLE AND SUBTITLE
Factor Structure of the Air Force Officer Qualifying Test Form S:
Analysis and Comparison with Previous Forms

5a. CONTRACT NUMBER
In-House

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S
Fritz Drasgow*
Christopher D. Nye*
Thomas R. Carretta**
Malcolm James Ree***

5d. PROJECT NUMBER
2313

5e. TASK NUMBER
HC

5f. WORK UNIT NUMBER
2313HC58

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Department of Psychology*
University of Illinois at Urbana-Champaign
603 E. Daniel St.
Champaign IL 61820

Our Lady of the Lake University***
San Antonio TX 78207

8. PERFORMING ORGANIZATION REPORT NUMBER
NA

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command**
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Supervisory Control Interfaces Branch
Wright-Patterson AFB OH 45433

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RHCI

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RH-WP-TR-2010-0008

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; Distribution is unlimited.

13. SUPPLEMENTARY NOTES

88 ABW PA Cleared 10/31/2008 , 88ABW-08-0866 .

14. ABSTRACT

Due to its importance for assignment and classification in the U. S. Air Force, the Air Force Officer Qualifying Test (AFOQT) has received a substantial amount of research. Recently, the AFOQT was revised to reduce administrative burden and test-taker fatigue. However, the new version, the AFOQT Form S, was implemented without explicitly examining the latent structure of the exam. The current study examined the factor structure of Form S as well as its measurement equivalence across race- and sex-based groups. Results indicated that a bifactor model with a general intelligence factor and five content-specific factors fit the best. The measurement equivalence of the AFOQT across gender and racial/ethnic groups also was supported.

15. SUBJECT TERMS
Air force Officer Qualifying Test, factor structure, confirmatory factor analysis

16. SECURITY CLASSIFICATION OF: **Unclassified**			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Thomas R. Carretta
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	29	**19b. TELEPONE NUMBER (*Include area code)***

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI-Std Z39-18

This page was intentionally left blank.

CONTENTS

TABLES

PREFACE

This report describes activities performed in support of USAF personnel selection and classification (AF/A1PF), Work Unit 2313HC58. The authors thank Mr Ken Schwartz (AETC/DPSF) and the AFPC Human Resources Data Bank (HRRD) for support in the development of the database used in this study.

INTRODUCTION

The *Air Force Officer Qualifying Test* (AFOQT) is the latest in a line of aptitude and achievement tests that traces its beginnings to the World War II Army Aviation Psychology Program (AAPP; Davis, 1947; Flanagan, 1948). Beginning in the early 1940s, many graduate students were offered commissions[1] to serve in the Army and conduct research on military and aviation related topics. The topics covered many aspects of assessment and classification into military specialties. The history of the program and its research is documented in a series of books published by the Government Printing Office in the late 1940s. The list of contributors reads like a who's who of psychometrics for the next four decades. Included are John Flanagan, Robert Thorndike, Lloyd Humphreys, Arthur Melton, Frederick B. Davis, and Philip DuBois (Flanagan, 1948).

Based on the AAPP results, the Army Air Corps and later the Air Force instituted apparatus-based tests such as the multidimensional pursuit test, drift correction test, stick and rudder manipulation, and complex coordination (Melton, 1947). There was a single centralized testing site where all apparatuses were maintained in careful calibration. When apparatus testing was decentralized to many sites it became nearly impossible to maintain administration consistency and the standards of calibration. The system became unworkable and was discontinued. It was proposed that paper-and-pencil tests replace the apparatus tests.

The immediate operational precursor of the AFOQT was the *Aviation Cadet Qualifying Test (*ACQT) which consisted of 13 subtests with a total of 300 items. Some of the content was expectable such as "General Mathematics" and "Mechanical Principles." Some was unexpected such as "Current Affairs" and "Biographical Data" heavily weighted to hunting and outdoor activities. The first AFOQT, Form A, was implemented in 1953 (Rogers, Roach, & Short, 1986; Valentine & Creager, 1961). Over the years, the test has gone through many versions and numerous modifications to its content. AFOQT Form A was issued with 665 items in 15 subtests. AFOQT Form B is notable as it was used for selection of the first class of the newly completed US Air Force Academy in Colorado Springs, Colorado. This form had 835 items in

[1] Frederick B. Davis told the story to my (Ree) class of graduate students in 1972 of how he was offered a commission or if he did not accept he could take his chances with the draft. He was commissioned in the US Army shortly thereafter and participated in the AAPP first in Santa Anna, California then in San Antonio, Texas.

17 subtests. Unique subtests included “Interests” and “Aerial Landmarks.” Form C was reduced to 645 items while Form D marked the apogee with 855 items with separate pilot and officer biographical inventories in 1957. Through Form G the AFOQT had close to 800 items. During the 1960s and 1970s there were small changes in the battery content and nomenclature and forms were called AFOQT-64, AFOQT-66, and AFOQT-68. Then there was a marked decline in the number of items on forms H through N. In 1978, Form N had 606 items in 18 subtests. With the implementation of Form O, four subtests (Background for Current Events, Tools, Aerial Landmarks, and Pilot Biographic and Attitude Scale) were removed from its immediate predecessor, Form N, and 2 new subtests were added (Aviation Information and Hidden Figures). AFOQT Forms O, P, Q, and R had 16 subtests with 380 items. Form R was never implemented but was eventually revised and published in 2005 as Form S with 250 items in 11 subtests.

The *Air Force Officer Qualifying Test* is used to award US Air Force (USAF) Reserve Officer Training Corps (ROTC) scholarships and to qualify applicants for officer commissioning through the ROTC and Officer Training School (OTS) programs. The AFOQT also is used to qualify applicants for aircrew training as pilots, combat system operators (formerly navigators), and air battle managers if they pass other educational, fitness, medical, moral, and physical requirements. For operational use, the subtests are combined into five overlapping composites (see Table 1). The Verbal, Quantitative, and Academic Aptitude composites are used to qualify applicants for ROTC and OTS officer commissioning programs. The Pilot and Navigator/Technical composites are used to qualify applicants for aircrew training. The AFOQT has been validated against officer training performance (Roberts & Skinner, 1996), several aircrew training performance criteria including passing/failing training, training grades, and class rank (Carretta, in press; Carretta & Ree, 2003; Olea & Ree, 1994), and several non-aviation officer jobs (Arth, 1986; Arth & Skinner, 1986; Finegold & Rogers, 1985; Hartke & Short, 1988).

Table 1. Composition of AFOQT Form S Aptitude Composites

	Composite				
Subtest	Verbal	Quantitative	Academic Aptitude	Pilot	Navigator/ Technical
Verbal Analogies (VA)	X		X		X
Arithmetic Reasoning (AR)		X	X	X	X
Word Knowledge (WK)	X		X		
Math Knowledge (MK)		X	X	X	X
Instrument Comprehension (IC)				X	
Block Counting (BC)					X
Table Reading (TR)				X	X
Aviation Information (AI)				X	
Rotated Blocks (RB)					
General Science (GS)					X
Hidden Figures (HF)					

Note. Although RB and HF were retained in AFOQT Form S, they do not contribute to any of the operational composite scores.

From 1980 through 2005, the AFOQT (Forms O, P, and Q) consisted of the same 16 subtests and each form was equated to the same normative score metric. Planned implementation of Form R was suspended as AFOQT program managers initiated a study to evaluate methods to reduce test administration time without adversely affecting its effectiveness. The goal was to determine the minimum test length or composition that maintained the current AFOQT psychometric characteristics. Successful achievement of these psychometric objectives would reduce the administration burden and examinee fatigue, and possibly make time available for new subtests with new content. Analyses indicated that five subtests could be removed while maintaining cognitive/knowledge content areas, reliability, and predictive validity, and avoiding an increase in adverse impact. Form R was revised and implemented as Form S in 2005. The administration time had been reduced from 4.5 to 3 hours with the removal of five subtests: Reading Comprehension, Data Interpretation, Mechanical Comprehension, Electrical Maze, and Scale Reading.

Form S was implemented without an empirical evaluation of its factor structure in a sample of applicants. Due to the substantial changes from earlier 16 subtest forms, the purpose of this study was to examine the latent factor structure of the 11 subtest AFOQT Form S and compare it to that of previous forms. In addition, the measurement equivalence of Form S was compared across gender and racial/ethnic groups. Measurement equivalence in increasingly important as the Air Force becomes more diverse.

Latent Structure of Earlier AFOQT Forms and Comparison to Form S

Carretta and Ree (1996) analyzed data from a sample of 3,000 applicants for Air Force commissions who had taken the 16 subtest AFOQT. Model 1 in their study corresponded to the operational composites in use at that time (but excluded Academic Aptitude because it was linearly dependent on the Verbal and Quantitative composites). This model was found to have a relatively poor fit to the data. Carretta and Ree's Model 2 was based on Skinner and Ree's (1987) exploratory factor analysis, which found verbal, math, spatial, aircrew, and perceptual speed factors. This model, with factors constrained to be orthogonal, also had a poor fit.

Carretta and Ree's (1996) Model 5 consisted of Model 2 augmented by a general factor, psychometric *g*, which enabled the model to account for correlations of subtests loading on different first order factors. This model provided an excellent fit to the data, with a root mean square error of approximation (RMSEA) of .071, a comparative fit index (CFI) of .957, and an average absolute standardized residual of .027. Because a model with several orthogonal content factors (i.e., verbal, math, etc.) and a single general factor is a nested submodel of a first order factor model with correlated factors (Yung, Thissen, & McLeod, 1999), Skinner and Ree's (1987) five factors (Model 2) would be expected to provide a good description of the AFOQT data if they were allowed to be oblique.

To compare the latent structure of the new Form S to earlier forms, two types of correlated factor models were fitted to the data. First, we fit models corresponding to the current AFOQT composites. This provides information about the degree to which the composites are aligned with the latent factors underlying the test battery. Second, we fit models based on Skinner and Ree's (1987) substantive factors. Because Carretta and Ree (1996) found these factors (with a general factor) to provide the best description of the 16 subtest AFOQT, we expected models based on this framework to fit well.

We also fit bifactor models to the AFOQT subtests. Although Carretta and Ree (1996) cited Schmid and Leiman (1957), their Model 5 is more closely related to a bifactor model than a Schmid-Leiman higher-order factor model. A bifactor model allows observed variables to load directly on a single general factor and a specific factor. For example, the Arithmetic Reasoning subtest would be expected to load on the general factor, a mathematical reasoning specific factor, and an error factor. Carretta and Ree's Model 5 has a bifactor structure with a few additional cross-loadings (e.g., the Block Counting subtest loaded on spatial and perceptual speed factors in addition to the general factor). A Schmid-Leiman higher-order model, in contrast is more restrictive than the bifactor model because it has additional proportionality constraints on factor loadings (see Yung et al., 1999, p. 115).

Form S of the AFOQT presents a challenge to confirmatory factor analysis because two of its composites and two of its factors are expected to have nonzero loadings for only two subtests. It is well known that statistical estimation of factor loadings requires at least three nonzero loadings. In this situation, a trick is sometimes used: one factor loading is fixed at a nonzero constant (e.g., one), one factor loading is estimated, and then the variance of that factor is treated as a free parameter to be estimated. It turns out that this approach does not actually estimate the factor loadings; it only estimates the ratio of the second loading to the first.

To better understand the latent structure of Form S, we also analyzed multi-item composites. We took this approach because the large number of items precluded an item-level factor analysis. Thus, for each subtest, mutually exclusive and exhaustive sets of items were used to form multi-item composites (called "item parcels" by Dorans & Lawrence, 1987) and then the composites were factor analyzed. For example, five composites were formed for the Verbal Analogies subtest and four composites were used for the Instrument Comprehension subtest. Because there were five parcels for the Arithmetic Reasoning and Math Knowledge subtests, factor loadings for 10 observed variables could be estimated for the mathematical reasoning factor and hence, factor loadings were statistically identified. Moreover, the sampling distribution of parcels more closely approximates the distribution assumed by linear factor analysis models (i.e., multivariate normality) than the sampling distribution of individual items.

Measurement Equivalence

The question addressed by measurement equivalence analyses is whether individuals with equal standings on the underlying trait assessed by a test, but sampled from different groups, have equal expected observed test scores (Drasgow, 1984). For example, do individuals with equal quantitative ability, but sampled from different groups, have equal expected scores on the Arithmetic Reasoning subtest? To examine measurement equivalence, mean and covariance structure (MACS; Sörbom, 1974) analysis was used. Here, the traditional factor analysis model,

$$x = \Lambda\xi + \delta,$$

is augmented to

$$x = \tau + \Lambda\xi + \delta,$$

where x is the vector of observed variables, Λ is the matrix of factor loadings, ξ is the vector of factor scores, and δ is a vector of errors. The difference between these two equations is the vector τ. Ordinarily, a correlation or covariance matrix is input to factor analysis; to this, MACS adds a vector of means of the observed variables and the vector τ contains the intercepts of the linear regressions of the observed variables x on the latent variables ξ.

To study measurement invariance across groups g, the MACS model is written as

$$x^{(g)} = \tau^{(g)} + \Lambda^{(g)}\xi^{(g)} + \delta^{(g)},$$

where the superscript g is used to indicate the group, $g = 1, \ldots, G$. Discussions of measurement invariance in the context of factor analysis (e.g., Ployhart & Oswald, 2004; Vandenberg & Lance, 2000) frequently mention configural, metric, and scalar invariance. Configural invariance is obtained when the pattern of fixed (at zero) and estimated factor loadings is identical across groups. Here the factor loadings may vary across groups, but each observed variable loads on the same factor(s) for all groups.

Metric invariance is a more restrictive condition than configural invariance. It imposes the constraint that the factor loading matrix, $\Lambda^{(g)}$ is identical across all the groups. Provided that configural invariance holds, metric invariance can be examined by the change in chi-square across configural and metric models: the metric invariance model is a nested submodel of the configural invariance model. Rather than a statistical test of the change in chi-square, ordinarily researchers examine the change in goodness of fit measures such as the RMSEA and the CFI.

Finally, scalar invariance is even more restrictive than metric invariance. Scalar invariance can be examined provided that metric invariance holds. Here, the constraint of invariant thresholds $\tau^{(g)}$ is added to invariant factor loadings $\Lambda^{(g)}$. Again, goodness of fit statistics should be examined to assess the extent to which adding this constraint degrades fit.

When scalar invariance is obtained, individuals with the same standings on the latent traits $\xi^{(g)}$ but sampled from different groups *g*, have the same expected observed score. In the language of item response theory (IRT), there is no differential item or test functioning. This is a very important property because it means that no group is disadvantaged by the test: one's underlying abilities ξ are transformed to observed scores in the same way for all groups.

It is important to note the distinction between measurement invariance and impact. In a MACS model, measurement invariance holds when $\tau^{(g)}$ and $\Lambda^{(g)}$ are invariant across groups. However, it is possible for groups to differ in their mean level of ability. For example, one group may have higher or lower means on the latent traits. We use the vector $\kappa^{(g)}$ to denote the factor means, $\kappa^{(g)} = E(\xi^{(g)})$ for group *g*. Thus, $\kappa^{(g)}$ can vary across groups because, without random assignment of people to groups there is no reason to expect groups to be equally skilled in the characteristics assessed by the subtests. Then invariant $\tau^{(g)}$ and $\Lambda^{(g)}$ mean that observed differences $E(x^{(g)})$ faithfully reflect the underlying differences on the factors.

In sum, we fit a variety of factor models to the AFOQT subtests and to multi-item composites formed from the subtests to examine the latent structure of this test battery. Then we examined measurement invariance across male/female and White/African American/Hispanic/Other groups to assess whether there was any evidence of differential item or test functioning.

METHODS

Sample

The data consisted of the responses of 12,511 applicants for USAF officer commissioning who were administered Form S1 of the AFOQT between 2005 and 2007. Mean age at time of testing was 22.4 years. The sample included 9,424 men (75%) and 2,978 women, 66%% were white, and more than 99% had completed at least a high school degree. In addition to qualification on the AFOQT, officer commissioning and aircrew training applicants met various academic (e.g., college degree), fitness (e.g., physical fitness test), moral (e.g., legal issues), medical (e.g., physical exam), and physical (e.g., weight) standards.

To assess equivalence across race and ethnicity, the sample was divided into four groups including White, African American, Hispanic, and Other. Because respondents were asked to indicate all of the races that applied to them, individuals marking more than one race were excluded from the analyses. This exclusion criterion resulted in samples of 8,296 Whites, 1,181 African Americans, 738 Hispanics, and 728 Others.

Measures

AFOQT Form S consists of 11 cognitive subtests that are combined into five composites. Personnel decisions including qualification for officer commissioning programs and aircrew training are made, in part, on the basis of the composites.

Brief descriptions of the AFOQT subtests grouped by content are presented below.

Verbal subtests. Verbal Analogies (VA) provides a measure of the ability to reason and determine relationships between words. Word Knowledge (WK) assesses verbal comprehension involving the ability to understand written language through the use of synonyms.

Quantitative subtests. Arithmetic Reasoning (AR) measures the ability to understand arithmetic relations expressed as word problems. Math Knowledge (MK) provides a measure of the ability to use mathematical terms, formulas, and relations.

Spatial subtests. Block Counting (BC) measures spatial ability through the analysis of three-dimensional representations of a set of blocks. Rotated Blocks (RB) assesses the ability to visualize and mentally manipulate objects. Hidden Figures (HF) measures the ability to see a simple figure embedded in a complex drawing.

Aircrew subtests. Instrument Comprehension (IC) assesses the ability to determine the attitude of an aircraft from illustrations of flight instruments. Aviation Information (AI) measures knowledge of general aviation terms, concepts, and principles. General Science (GS) provides a measure of knowledge and understanding of scientific, terms, concepts, instruments, and principles.

Perceptual speed subtest. Table Reading (TR) assesses the ability to quickly and accurately extract information from tables.

Procedures

Our starting model was based on a confirmatory model of the previous 16 subtest version of the AFOQT (Carretta & Ree, 1996). This model consisted of a factor representing general cognitive ability (*g*) and five specific cognitive factors of verbal, math, spatial, aircrew knowledge, and perceptual speed.

Analyses

Several goodness-of-fit statistics were considered. Our choice of fit indices was guided in part by Hu and Bentler (1998, 1999) who recommend using both an incremental fit index and an absolute fit index to examine model fit. We chose the incremental fit indices of the Goodness-of-Fit Index (GFI; Tanaka & Huba, 1985), the Adjusted Goodness-of-fit Index (AGFI; Jöreskog & Sörbom, 1989), the Comparative Fit Index (CFI; Bentler, 1990, 1995), and the Non-Normed Fit Index (NNFI; Bentler & Bonnett, 1980). The absolute fit indices we examined were the Standardized Root Mean Square Residual (Hu & Bentler, 1999) and the Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993). Hu and Bentler (1999) recommended the following cutoff values as indicators of good model fit: NNFI and CFI of .95 or higher, SRMR of .08 or less, and RMSEA of .06 or less. In addition, previous research has suggested that a GFI of .95 (Marsh & Grayson, 1995) and an AGFI of .90 (Schermelleh-Engel, Moosbrugger, & Muller, 2003) reflect acceptable model fit.

After estimating the CFA, eigenvalue and eigenvector analyses were conducted to compare the Form S Pilot and Navigator-Technical composites with the same named composites from a previous AFOQT form with 16 subtests. AFOQT Forms O, P, and Q had the same 16 subtests and were equated to a common scale. The goal was to assess the first factor saturation of each composite and to identify the relative contribution of *constructs* to the composites. Form O data were used because previous CFAs of the 16 subtest AFOQT were accomplished using Form O data (Carretta & Ree, 1996).

RESULTS

Table 2 shows the correlation matrix for the 11 subtests on Form S1 (the correlation matrix for the item parcel matrix is available from the first author). All correlations in Table 2 are positive. The correlations range from .182 (WK and TR) to .706 (AR and MK) with a mean of .413. These values are similar to those reported for the 16 subtest version where the correlations ranged from .17 (WK and EM) to .77 (RC and WK) with a mean of .436 (Carretta & Ree, 1996). An eigenvalue analysis of the adjusted correlation matrix (principal axis factoring) showed general cognitive ability, *g*, accounted for 47 percent of the variance. This was estimated from the first unrotated principal factor as discussed by Ree and Earles (1991).

Table 2. AFOQT Form S Subtest Correlation Matrix

Subtest	VA	AR	WK	MK	IC	BC	TR	AI	GS	RB	HF
VA	1.000										
AR	0.514	1.000									
WK	0.691	0.446	1.000								
MK	0.422	0.706	0.358	1.000							
IC	0.376	0.456	0.310	0.400	1.000						
BC	0.350	0.484	0.271	0.410	0.508	1.000					
TR	0.249	0.389	0.182	0.309	0.332	0.485	1.000				
AI	0.371	0.349	0.363	0.282	0.612	0.357	0.250	1.000			
GS	0.561	0.530	0.548	0.566	0.471	0.361	0.210	0.485	1.000		
RB	0.353	0.478	0.284	0.423	0.563	0.507	0.306	0.418	0.443	1.000	
HF	0.348	0.422	0.270	0.396	0.485	0.492	0.335	0.333	0.377	0.543	1.000

Notes. N = 12, 511; All correlations were significant at the p≤ .01 level of significance.

Confirmatory Factor Analysis

The fit statistics for the models we examined are shown in Table 3. As shown, the single factor model fit the data poorly: analysis of the eleven subtests produced an RMSEA of .17 and analysis of multi-item composites (i.e., parcels) yielded an RMSEA of .15. Both of these RMSEAs are well above the range that would be considered a good fit (Hu & Bentler, 1999). Although the four factor model corresponding to the operational composites fit reasonably well, the bifactor solution with five specific factors fit better. Specifically, the model with a general factor and five content factors (verbal, math, spatial, aircrew, and perceptual speed) had an RMSEA of .059, a NNFI of .97, a CFI of .97, and an SRMR of .044 when analyzing the parcels. Interestingly, the GFI and AGFI indices were noticeably lower for all of the models involving parcels, even when all of the other fit statistics indicated an excellent fit. Our experience is that GFI and AGFI are sensitive to model complexity: They tend to be lower in models with more manifest variables. Thus, we believe that the observed values of GFI and AGFI indicate satisfactory fits for the five-factor model with correlated factors and the bifactor model with five specific factors. In sum, similar to Carretta and Ree's (1996) Model 5, our results indicate that the data is best represented by a general intelligence factor and five content-specific factors (verbal, quantitative, spatial, aircrew, and perceptual speed).

Table 3. Fit statistics for Confirmatory Factor Analysis Models

Model	RMSEA	NNFI	GFI	SRMR	GFI	AGFI
Single Factor Model - Parcels	0.150	0.89	0.89	0.10	0.49	0.45
Single Factor Model- Subtests	0.170	0.85	0.88	0.082	0.80	0.71
Four-Factor (Composites) Model – Parcels	0.075	0.96	0.96	0.062	0.79	0.77
Four-Factor Model – Parcels (Phi = Id)	0.086	0.94	0.94	0.24	0.74	0.71
Bifactor with Composite Specific Factors – Parcels	0.065	0.97	0.97	0.05	0.84	0.82
Five-Factor (V, M, Sp, AC, PS) Model – Subtests	0.078	0.97	0.98	0.033	0.97	0.93
Five-factor (V, M, Sp, AC, PS) – Parcels	0.059	0.97	0.97	0.044	0.86	0.84
Five-Factor Model – Parcels (Phi = Id)	0.072	0.96	0.96	0.22	0.80	0.79
Bifactor with Composite Specific Factors – Parcels	0.053	0.98	0.98	0.057	0.88	0.87

Note. Phi = Id indicates that the factor correlation matrix Φ was restricted to an identity matrix, i.e., factors were orthogonal.

The parameter estimates of the bifactor model are also informative. Although all indicators had substantial loadings on both the general and their specific factors, the majority had their strongest loading on a specific dimension. Notable exceptions included the Block Counting, Rotated Blocks, and Hidden Figures subtests, which had their highest loadings on the general factor. This suggests that a large portion of the variance in the general factor is accounted for by spatial ability. Nonetheless, most indicators still had substantial loadings on the general factor with estimates ranging from .28 (WK2) to .70 (BC3).

We analyzed the data with bifactor models where each observed variable loaded on the general factor and one content factor and, to enhance the comparability of our results to those of Carretta and Ree (1996), we obtained solutions where the Block Counting subtest was allowed to load on the general, spatial, and perceptual speed factors and the General Science subtest was allowed to load on the general, verbal, and aircrew factors. Results were very similar for these two types of models and, consequently, Table 3 presents fit statistics for the analyses parallel to Carretta and Ree.

When we allowed cross-loadings, the Block Counting subtest had negative loadings on the spatial dimension and positive loadings on the perceptual speed dimension. The negative loadings may be a result of the magnitude of this subtest’s relationship with the general factor. Factor loadings for this model are shown in Table 4.

Table 4. Completely Standardized Solution for the Bifactor Model with Five Specific Factors

Parcel	General	Verbal	Quantitative	Spatial	Aircrew	Perceptual Speed
VA 1	.39	.46	--	--	--	--
VA 2	.31	.48	--	--	--	--
VA 3	.33	.56	--	--	--	--
VA 4	.33	.31	--	--	--	--
VA 5	.39	.53	--	--	--	--
AR 1	.52	--	.46	--	--	--
AR 2	.52	--	.49	--	--	--
AR 3	.50	--	.51	--	--	--
AR 4	.49	--	.52	--	--	--
AR 5	.47	--	.46	--	--	--
WK 1	.29	.59	--	--	--	--
WK 2	.28	.61	--	--	--	--
WK 3	.30	.67	--	--	--	--
WK 4	.31	.70	--	--	--	--
WK 5	.33	.70	--	--	--	--
MK 1	.39	--	.60	--	--	--
MK 2	.45	--	.57	--	--	--
MK 3	.40	--	.52	--	--	--
MK 4	.45	--	.57	--	--	--
MK 5	.48	--	.56	--	--	--
IC 1	.59	--	--	--	.58	--
IC 2	.56	--	--	--	.57	--
IC 3	.62	--	--	--	.58	--
IC 4	.58	--	--	--	.53	--
BC 1	.69	--	--	-.33	--	.06
BC 2	.66	--	--	-.28	--	.09
BC 3	.70	--	--	-.31	--	.11
BC 4	.67	--	--	-.29	--	.12
TR 1	.37	--	--	--	--	.67
TR 2	.39	--	--	--	--	.65
TR 3	.40	--	--	--	--	.72
TR 4	.38	--	--	--	--	.69
TR 5	.39	--	--	--	--	.66
TR 6	.34	--	--	--	--	.65
TR 7	.36	--	--	--	--	.70

Table 4. Completely Standardized Solution for the Bifactor Model with Five Specific Factors, continued

Parcel	General	Verbal	Quantitative	Spatial	Aircrew	Perceptual Speed
TR 8	.36	--	--	--	--	.71
AI 1	.42	--	--	--	.47	--
AI 2	.34	--	--	--	.49	--
AI 3	.43	--	--	--	.47	--
AI 4	.34	--	--	--	.48	--
GS 1	.41	.32	--	--	.08	--
GS 2	.33	.36	--	--	.13	--
GS 3	.45	.35	--	--	.19	--
GS 4	.41	.30	--	--	.20	--
RB 1	.65	--	--	.09	--	--
RB 2	.61	--	--	.08	--	--
RB 3	.57	--	--	.08	--	--
HF 1	.66	--	--	.41	--	--
HF 2	.68	--	--	.40	--	--
HF 3	.67	--	--	.37	--	--

Measurement Invariance

Carretta and Ree (1995) investigated the invariance of factor loadings for an earlier form of the 16 subtest AFOQT using a sample of 269,968 applicants for U. S. Air Force commissions that were tested between 1981 and 1993. They compared males (N = 219,887) and females (N = 50,081) and also compared Black (N = 32,798), Hispanic (N = 12,647), Asian-American (N = 9,460), and Native-American (N=2,551) groups to Whites (N = 212,238). Given these very large sample sizes, it is not surprising that they found statistically significant differences in factor loadings. More importantly, however, they found that the differences in factor loading were very small in size (generally less than .05), indicating that the tests functioned equivalently across groups.

Following Carretta and Ree (1995), we examined the measurement equivalence of the bifactor model with five content factors because it fit the best in the total sample. We also tested the equivalence of the four factor model because of its operational use by the Air Force. Table 5 shows the results of the MACS analyses using a factor pattern matrix based on the current operational composites with correlated factors and Table 6 gives the results for the bifactor structure with a general factor and the five content factors described above. Similar to Carretta and Ree (1996), the Block Counting and General Science subtests were allowed to cross-load on additional specific factors in the bifactor model. Both Tables 5 and 6 show only negligible changes in the fit indices when the constrained models are compared to the baseline (i.e., configural) solution.

Table 5. Fit Statistics for the MACS Analyses of the Four-Factor Structure

				SRMR	
Model (Sex)	RMSEA	NNFI	CFI	Male	Female
Configural Invariance	0.074	0.95	0.96	0.068	0.056
Metric Invariance	0.074	0.95	0.96	0.067	0.071
Scalar Invariance	0.075	0.95	0.95	0.068	0.086

				SRMR			
Model (Race)	RMSEA	NNFI	CFI	White	African American	Hispanic	Other
Configural Invariance	0.075	0.95	0.95	0.068	0.061	0.069	0.070
Metric Invariance	0.073	0.95	0.95	0.068	0.076	0.080	0.077
Scalar Invariance	0.074	0.95	0.95	0.066	0.086	0.086	0.081

Table 6. Fit Statistics for the MACS Analyses of the Bifactor Structure

				SRMR	
Model (Sex)	RMSEA	NNFI	CFI	Male	Female
Configural Invariance	0.052	0.98	0.98	0.061	0.053
Metric Invariance	0.051	0.98	0.98	0.062	0.066
Scalar Invariance	0.053	0.97	0.98	0.063	0.078

				SRMR			
Model (Race)	RMSEA	NNFI	CFI	White	African American	Hispanic	Other
Configural Invariance	0.052	0.97	0.97	0.059	0.054	0.062	0.068
Metric Invariance	0.051	0.97	0.97	0.060	0.063	0.074	0.098
Scalar Invariance	0.053	0.97	0.97	0.060	0.075	0.090	0.120

In the analysis of measurement invariance for race, RMSEA did not change whatsoever from the configural invariance model to the scalar invariance model: It was .054 for both, as shown in Table 6 for the model with a general factor and five content factors. The NNFI and CFI measures also showed no change. For the individual groups, moderate changes in SRMRs were observed for African Americans, Hispanics, and Others, probably because these samples were much smaller than the White sample.

A similar pattern of results is apparent for the male-female comparison. Table 6 shows that the RMSEA had a trivial increase (.054 to .055), and the NNFI and CFI did not change at all. The male SRMR did not change, and the female SRMR increased moderately, probably because this sample was much smaller than the male sample. In sum, the results suggest that there is little or no differential item and test functioning across minority and majority groups.

Comparison of Form S Pilot and Navigator/Technical Composites with Previous Forms

Eigenvalue and eigenvector analyses of the AFOQT Form S Pilot composite showed only one eigenvalue over 1.0 accounting for 53% of the variance. A similar result was found for the Navigator/Technical composite with only one eigenvalue over 1.0 accounting for 55% of the variance.

The same analyses for the previous AFOQT Form O showed that the first eigenvalue for the Pilot composite accounted for 50% of the variance in the matrix and a second eigenvalue over 1.0 accounted for 13%. The Form O Navigator/Technical composite had one eigenvalue over 1.0 that accounted for 52% of the variance.

DISCUSSION

In this study, a variety of confirmatory factor analysis models were fit to data from the recently revised AFOQT. The results and conclusions are strikingly similar to those obtained by Carretta and Ree (1996): an important general cognitive ability factor underlies performance on all of the subtests and verbal, mathematical reasoning, spatial, aircrew, and perceptual speed factors underlie groups of subtests. Moreover, excellent fits were obtained, so we can have confidence in these findings.

Mean and covariance structure analysis was used to investigate measurement invariance for the AFOQT. Very positive results were obtained in that the overall fit statistics for the most restrictive models (i.e., models specifying scalar invariance) were nearly the same size as the fit statistics for the least restrictive models (i.e., models specifying configural invariance). This indicates that AFOQT scores can be used to make comparisons across candidates irrespective of their gender or race.

Operationally, personnel measurement, selection and classification decisions involving the AFOQT are based on composite scores. The Pilot and Navigator/Technical composites are part of the system for aircrew training qualification including pilots, combat system operators, and air battle managers. Therefore, the nature and performance of these composites is very important. Comparison of the prior Forms O, P, and Q and current Form S composite scores and underlying structure is informative.

On the first eigenvector for Form S, Arithmetic Reasoning had the greatest loading (largest value eigenvector) at .502 and Table Reading showed the smallest at .362. The results for the first factor from the Form O Pilot composite showed high eigenvector values for the perceptual speed, spatial, and aviation job knowledge subtests of Scale Reading (.795), Block Counting (.785), Mechanical Comprehension (.750), and Instrument Comprehension (.742). The magnitudes of the Form S loadings are much smaller than the loadings for Form O. Further, Form S has subtests that are more indicative of *g* than Form O. The subtests on the Form O Pilot composite all share the characteristic that the male means are noticeably greater than the female means. This is not the case in Form S. The newly implemented Form S should be expected to have smaller male-female differences than Form O.

For the Form S Navigator-Technical composite, the first factor accounted for 55% of the variance with Arithmetic Reasoning again showing the greatest loading at .512 and Table Reading the lowest at .304. High-level findings for Form O were similar. The first factor accounted for 52% of the variance. The highest loading on the first factor was Scale Reading (.815) followed by the three mathematics tests; Arithmetic Reasoning (.807), Data Interpretation (.766), and Math Knowledge (.788). Electrical Maze (.612) showed the lowest value. Mechanical Comprehension, Data Interpretation, Scale Reading, and Electrical Maze were all removed from Form S. Further, Rotated Blocks and Hidden Figures were removed from the Navigator/Technical composite. The Form O Navigator/Technical composite is composed of four subtests that are good indicators of *g*, Verbal Analogies, Arithmetic Reasoning, Math Knowledge, and Block Counting. It also includes a marker for perceptual speed in Table Reading, the only speeded subtest on the AFOQT, and General Science which has loadings on verbal and aircrew factors. On both Form S and Form O there was not a second eigenvalue equal to or above 1 for the Navigator/Technical composite. Given the change in the content of Forms S a difference in validity might be expected. However, the presence of the highly *g* loaded subtests suggests otherwise.

In a series of papers, Ree and colleagues (Olea & Ree, 1994; Ree, Carretta, & Teachout, 1996; Ree, & Earles, 1991; Ree, Earles, & Teachout, 1994) showed that psychometric *g* was largely responsible for predicting performance in training and on the job for a wide variety of military samples. It appears that the current form of the AFOQT taps psychometric *g* and would be expected to predict performance in ways similar to previous forms.

REFERENCES

Arth, T.O. (1986). *Validation of the AFOQT for non-rated officers,* AFHRL-TP-85-50. Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Arth, T. O., Skinner, J. (1986). *Aptitude selection for Air Force officer non-aircrew jobs.* Paper presented at the annual meeting of the Military Testing Association, Mystic, CT.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software, Inc.

Browne, M. W., Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Lang (Eds.), *Testing structural equation models.* Newbury Park, CA: Sage.

Carretta, T. R. (in press). *Predictive validity of the Air Force Officer Qualifying Test for USAF air battle manager training performance,* AFRL-RH-WP-TR-2008-xxxx. Wright-Patterson AFB, OH: Air Force Research Laboratory, Warfighter Interfaces Division.

Carretta, T. R., Ree, M. J. (1995). Near identity of cognitive structure in sex and ethnic groups. *Personality and Individual Differences, 19,* 149-155.

Carretta, T. R., Ree, M. J. (1996). Factor structure of the Air Force Officer Qualifying Test: Analysis and comparison. *Military Psychology, 8,* 29-42.

Carretta, T. R., Ree, M. J. (2003). Pilot selection methods. In B. H. Kantowitz (Series Ed.) & P. S. Tsang & M. A. Vidulich (Vol. Eds.). *Human factors in transportation: Principles and practices of aviation psychology* (pp. 357-396). Mahwah, NJ: Erlbaum.

Davis, F. B. (1947). *The AAF qualifying examination,* Report No. 6. Washington, DC: U.S. Government Printing Office.

Dorans, N. J., Lawrence, I. M. (1987). *The internal construct validity of the SAT* (RR-87-35). Princeton, NJ: Educational Testing Service.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95*, 134-135.

Finegold, L., Rogers, D. (1985). *Relationship between Air Force Officer Qualifying Test scores and success in air weapons controller training,* AFHRL-TR-85-13.Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Flanagan, J. C. (1948). *The aviation psychology program in the Army Air Forces.* Report No. 1. Washington, DC: U.S. Government Printing Office.

Hartke, D. D., Short. L. O. (1988). *Validity of the academic aptitude composite of the Air Force Officer Qualifying Test (AFOQT),* AFHRL-TP-87-61. Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fir indices. *Sociological Methods of Research, 11,* 325-344.

Hu, L. T., Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424-453.

Hu, L. T.; Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Jöreskog, K. G., Sörbom, D. (1993). *LISREL 8 user's reference guide.* Chicago: Scientific Software International.

Marsh, H. W., Balla, J. R., McDonald, R. P. (1998). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391-410.

Marsh, H. W., Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Issues, concepts, and applications* (pp. 56-75). Newbury Park, CA: Sage.

Melton, A. W. (1947). Apparatus tests, Report No. 4. Washington, DC: U.S. Government Printing Office.

Olea, M. M., Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology, 79,* 845-849.

Ployhart, R. E., Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods, 7,* 27-65.

Ree, M. J., Carretta, T. R., Teachout, M. S. (1995). Role of ability and prior job knowledge in complex training performance. *Journal of Applied Psychology, 80,* 721-730.

Ree, M. J., Earles, J. A. (1991). The stability of convergent estimates of *g*. *Intelligence, 15*, 271-278.

Ree, M. J., Earles, J. A., Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79*, 518-524.

Roberts, H. E., & Skinner, J. (1996). Gender and racial equity of the Air Force Officer Qualifying Test in officer training school selection decisions. *Military Psychology, 8,* 95-113.

Rogers, D. L., Roach, B. W., Short, L. O. (1986). *Mental ability testing in the selection of Air Force officers: A brief historical overview,* AFHRL-TP-86-23. Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Schermelleh-Engel, K., Moosbrugger, H., Muller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*, 23-74.

Schmid, J., Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61.

Skinner, J., Ree, M. J. (1987). *Air Force Officer Qualifying Test (AFOQT): Item and factor analysis of Form O,* AFHRL-TR-86-68. Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.

Tanaka, J. S., Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 42*, 233-239.

Valentine, L. D. Jr., Creager, J. A. (1961). *Officer selection and classification tests: Their development and use,* ASD-TN-61-145. Lackland AFB, TX: Personnel Laboratory, Aeronautical Systems Division.

Vandenberg, R. J., Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 2*, 4-69.

Yung, Y.-F., Thissen, D., McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*, 113-128